Prediction and Evaluation with Bradley-Terry: A College Hockey Case Study

John T. Whelan
jtwsma@rit.edu

School of Mathematical Sciences
and Center for Computational Relativity & Gravitation
Rochester Institute of Technology

Presented at UP-STAT 2018
University of Rochester 2018 April 21

Outline

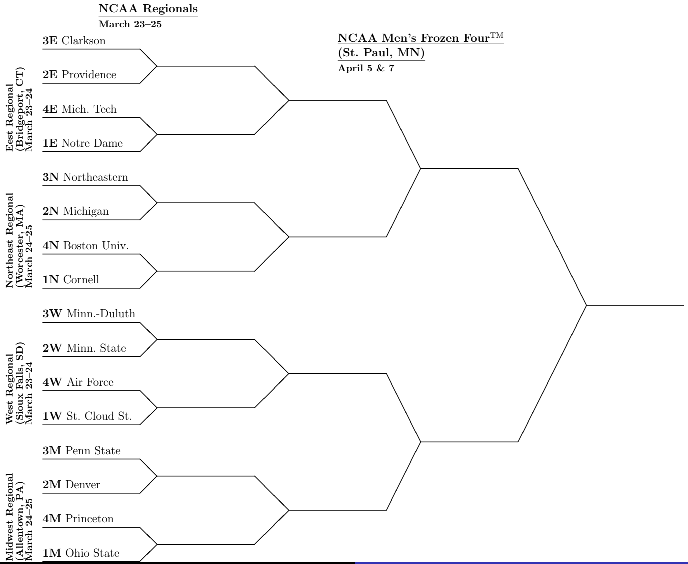
Outline

1. Evaluation of Predictions Using the Bayes Factor

2. Bayesian Bradley-Terry Models
   - Haldane (Maximum Likelihood)
   - Beta (Generalized Logistic)
   - Gaussian

3. Hierarchical Bayesian Bradley-Terry Models

NCAA Hockey Tournament



Example of Probabilistic Predictions

- Model ☞ predicted probs for outcome of each (possible) game
  Based on info including regular season results

Cr def BU	0.683	BU def Cr	0.317
Mi def NE	0.541	NE def Mi	0.459
Cr def Mi	0.624	Mi def Cr	0.376
Cr def NE	0.661	NE def Cr	0.339
BU def Mi	0.436	Mi def BU	0.564
BU def NE	0.476	NE def BU	0.524

- Note: outcomes may be correlated
  e.g., team strength not well constrained by model
- Combine ☞ prob for each of 2^{15} possible "bracket" outcomes

Assessing Predictions with the Bayes Factor

- Odds ratio for model comparison

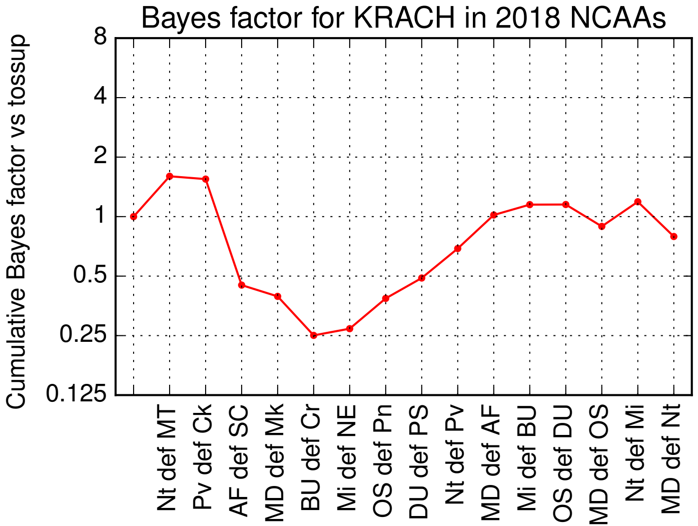
$$\frac{P(\mathcal{M}_1|D,I)}{P(\mathcal{M}_2|D,I)} = \frac{P(D|\mathcal{M}_1,I)}{P(D|\mathcal{M}_2,I)} \frac{P(\mathcal{M}_1|I)}{P(\mathcal{M}_2|I)}$$

- Here $D \equiv$ tourney results
  Bayes factor $\frac{P(D|\mathcal{M}_1,I)}{P(D|\mathcal{M}_2,I)}$ only needs prob of actual set of results (not all 2^{15} possible)
- Convenient to compare each model to “tossup” model where outcome of each game is 50-50:

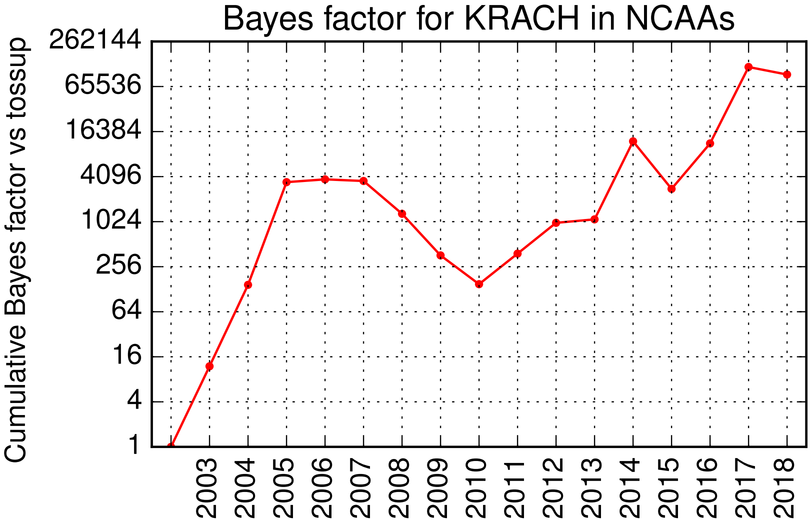
$$\mathcal{B}(\mathcal{M}) = \prod_{\text{game}} 2 \times P(\text{winner}|\mathcal{M}, I)$$

For each game, pick up a factor between 0 and 2.

Bayes Factor Example



Cumulative Bayes Factor, 2003-2018



Outline

KRACH = Maximum Likelihood Bradley-Terry

- Model: each team $i \in \{1, \ldots, t\}$ has log-strength[1] λ_i

$$P(i \text{ def } j) = \theta_{ij} = \frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j}} = \text{logistic}(\lambda_i - \lambda_j)$$

Zermelo 1929, Bradley & Terry 1952, (Butler 1993)

- Given (reg season) results: w_{ij} wins in n_{ij} games for i against j, KRACH ratings $\{\widehat{\lambda}_i\}$ satisfy

$$v_i = \sum_{j=1}^{t} w_{ij} = \frac{e^{\widehat{\lambda}_i}}{e^{\widehat{\lambda}_i} + e^{\widehat{\lambda}_j}} = \sum_{j=1}^{t} n_{ij} \widehat{\theta}_{ij}$$

and maximize likelihood

$$p(\{w_{ij}\} | \{\lambda_i\}, \{n_{ij}\}) = \prod_{i=1}^{t} \prod_{j=1}^{t} \binom{n_{ij}}{w_{ij}} \theta_{ij}{}^{w_{ij}}$$

[1] or a strength $\pi_i = e^{\lambda_i}$

Bradley-Terry w/Haldane Prior

- ML $\{\widehat{\lambda}_i\} \equiv$ maximum a posteriori w/uniform (improper) "Haldane" prior[2] $f(\{\lambda_i\}|I_0) = \text{const}$
- Bayesian approach would marginalize over posterior

$$f(\{\lambda_i\}|w_{ij}, I_0) \propto p(\{w_{ij}\}|\{\lambda_i\}, \{n_{ij}\})$$

to get $P(D|w_{ij}, I_0) = \int d^t\lambda\, P(D|\{\lambda_i\}, I_0)\, f(\{\lambda_i\}|w_{ij}, I_0)$

- Approximate with Gaussian expansion about MAP point

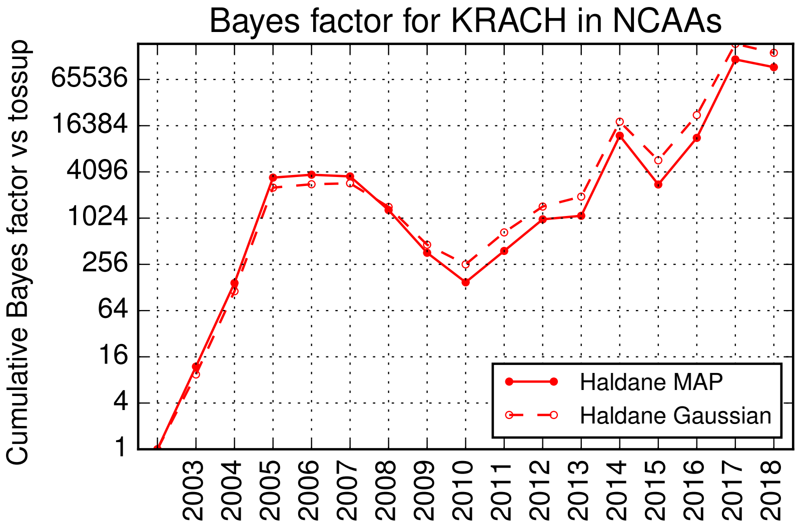
$$f_G(\{\lambda_i\}|w_{ij}, I_0) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{t}\sum_{j=1}^{t}\left(\lambda_i - \widehat{\lambda}_i\right) H_{ij}\left(\lambda_j - \widehat{\lambda}_j\right)\right)$$

with inverse variance-covariance matrix

$$H_{ij} = -\left.\frac{\partial^2 \ln f(\{\lambda_i\}|w_{ij}, I_0)}{\partial\lambda_i\,\partial\lambda_j}\right|_{\{\lambda_i=\widehat{\lambda}_i\}} = -n_{ij}\widehat{\theta}_{ij}\widehat{\theta}_{ji} + \delta_{ij}\sum_{k=1}^{t} n_{ik}\widehat{\theta}_{ik}\widehat{\theta}_{ki}$$

[2] by analogy to Haldane 1932 in binomial case

Cumulative Bayes Factor, 2003-2018



Bradley-Terry w/Beta Prior

- Haldane prior improper ☞ extreme results w/e.g., undefeated teams
  Prefer proper prior which satisfies desiderata of Whelan 2018, e.g., Beta(η,η) in $\zeta_i = \text{logistic}(\lambda_i) = \frac{\pi_i}{1+\pi_i} \in (0,1)$:

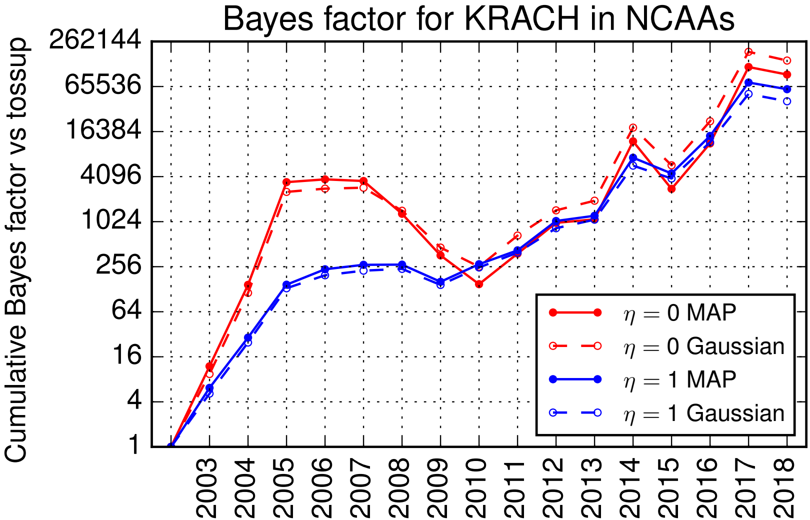
$$f(\lambda_i|I_\eta) \propto \left(1+e^{-\lambda_i}\right)^{-\eta}\left(1+e^{\lambda_i}\right)^{-\eta}$$

  $\eta \to 0$ is Haldane; $\eta = 1$ is uniform in ζ_i

- MAP equations: $v_i + \eta = \sum_{j=1}^{t} n_{ij}\widehat{\theta}_{ij} + 2\eta\widehat{\zeta}_i$
  MLE w/2η games "split" vs "fictitious team" w/log-strength 0
  $\eta = \frac{1}{2}$ sometimes used to regularize KRACH (Butler 1993)
- Gaussian approx w/

$$H_{ij} = -n_{ij}\widehat{\theta}_{ij}\widehat{\theta}_{ji} + \delta_{ij}\left(\sum_{k=1}^{t} n_{ik}\widehat{\theta}_{ik}\widehat{\theta}_{ki} + 2\eta\widehat{\zeta}_i(1-\widehat{\zeta}_i)\right)$$

Cumulative Bayes Factor, 2003-2018



Bradley-Terry w/Gaussian Prior

- Convenient to work with Gaussian prior on $\{\lambda_i\}$ (Leonard 1977; Whelan 2018)

$$f(\lambda_i|I_\sigma) \propto \exp\left(-\frac{{\lambda_i}^2}{2\sigma^2}\right)$$

$\sigma \to \infty$ is Haldane

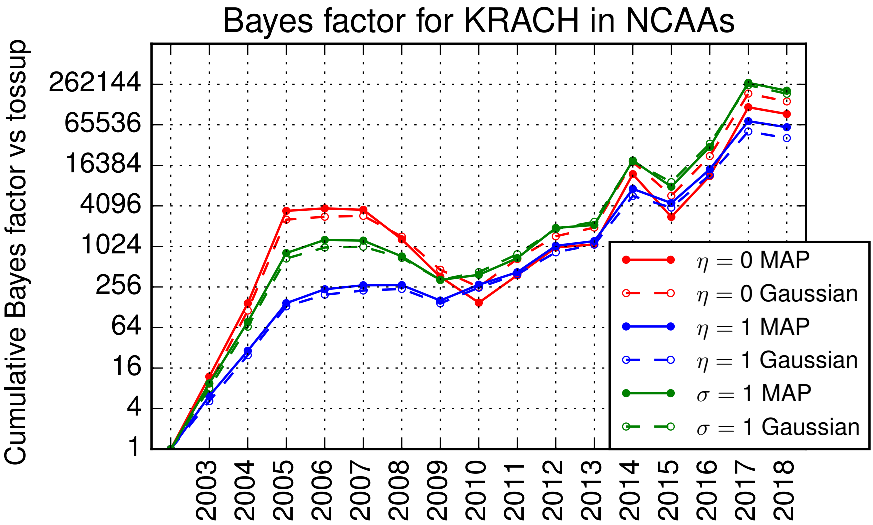
- MAP equations:

$$v_i = \sum_{j=1}^{t} n_{ij}\widehat{\theta}_{ij} + \frac{\widehat{\lambda}_i}{\sigma^2}$$

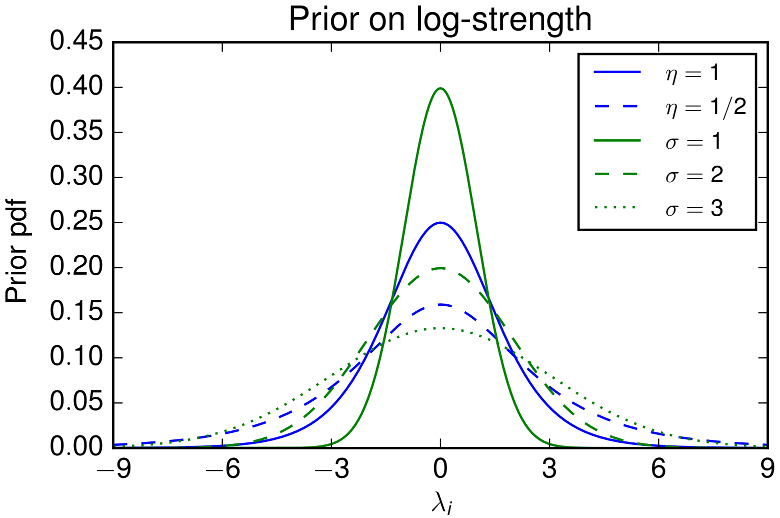
- Gaussian approx w/

$$H_{ij} = -n_{ij}\widehat{\theta}_{ij}\widehat{\theta}_{ji} + \delta_{ij}\left(\sum_{k=1}^{t} n_{ik}\widehat{\theta}_{ik}\widehat{\theta}_{ki} + \frac{1}{\sigma^2}\right)$$

Cumulative Bayes Factor, 2003-2018



Comparison of Prior Distributions



Outline

1. Evaluation of Predictions Using the Bayes Factor
2. Bayesian Bradley-Terry Models
    - Haldane (Maximum Likelihood)
    - Beta (Generalized Logistic)
    - Gaussian
3. Hierarchical Bayesian Bradley-Terry Models

Motivation for Hierarchical Model

- Values of η in I_η & σ in I_σ arbitrary
- Use hierarchical model w/hyperprior on hyperparameter (Phelan & Whelan 2018) e.g.,

$$f(\{\lambda_i\},\sigma|I_H) \propto \sigma^{-t} \exp\left(-\frac{\sum_{i=1}^t {\lambda_i}^2}{2\sigma^2}\right) f(\sigma|I_H)$$

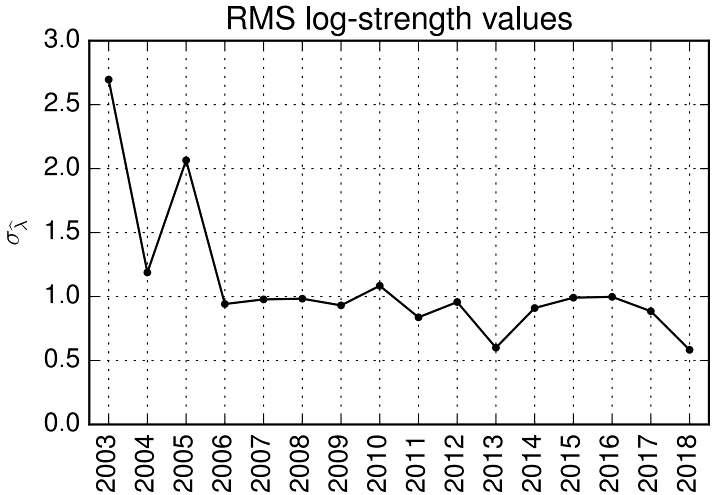
- Uniform hyperprior on σ?

  MAP eqns $v_i = \sum_{j=1}^{t} n_{ij}\widehat{\theta}_{ij} + \frac{\widehat{\lambda}_i}{\widehat{\sigma}^2}$ and $\widehat{\sigma}^2 = \frac{1}{t}\sum_{i=1}^t {\widehat{\lambda}_i}^2$

  Problem: MAP point is at $\sigma \to 0$
- Phelan & Whelan 2018 used $\Gamma(\alpha,\beta)$ prior motivated by variance of MLE $\{\widehat{\lambda}_i\}$ from previous season

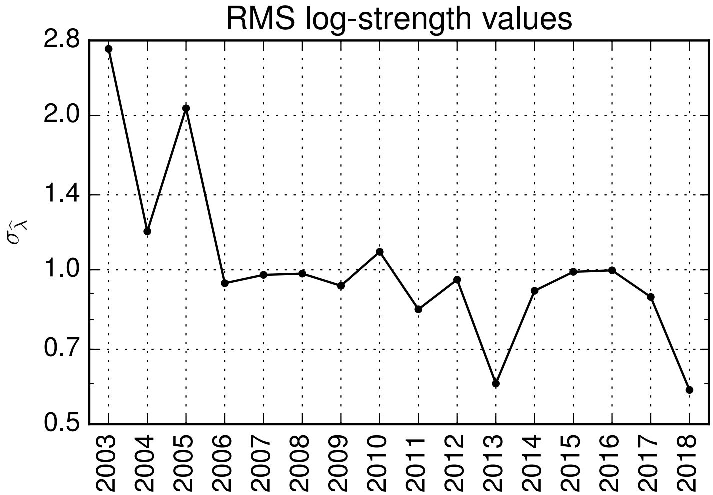
  Still have $\widehat{\sigma} \to 0$ unless $\alpha \geq t+1$

Standard Deviation of ML Log-Strengths



Compare $\sigma \approx 0.27$ for MLB (Phelan & Whelan 2018)

Standard Deviation of ML Log-Strengths



mean($\ln \sigma$) $= 0.019$, std($\ln \sigma$) $= 0.379$

Log-Normal Hierarchical Model

- Use log-normal prior on σ w/empirical parameters $\ln\sigma_0$ & ε:

$$f(\{\lambda_i\}, \ln\sigma | I_L) \propto \sigma^{-t} \exp\left(-\frac{\sum_{i=1}^{t}\lambda_i{}^2}{2\sigma^2} - \frac{(\ln\sigma - \ln\sigma_0)^2}{2\varepsilon^2}\right)$$

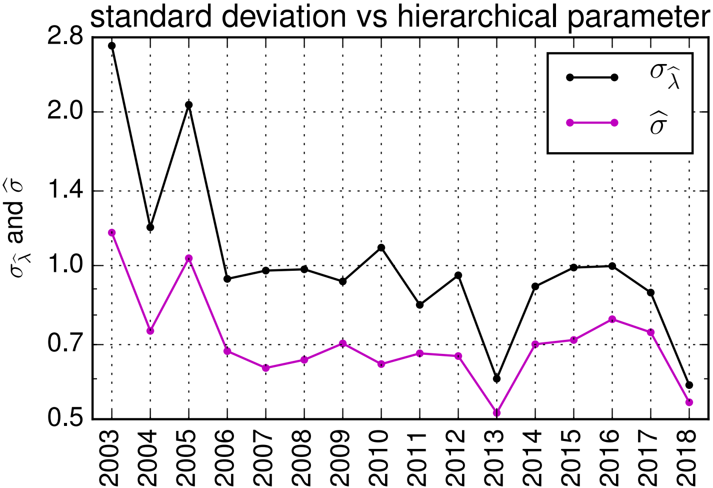
- MAP eqns

$$v_i = \sum_{j=1}^{t} n_{ij}\widehat{\theta}_{ij} + \frac{\widehat{\lambda}_i}{\widehat{\sigma}^2} \qquad \text{and} \qquad t = \frac{\sum_{i=1}^{t}\widehat{\lambda}_i{}^2}{\widehat{\sigma}^2} - \frac{\ln\widehat{\sigma} - \ln\sigma_0}{2\varepsilon^2}$$

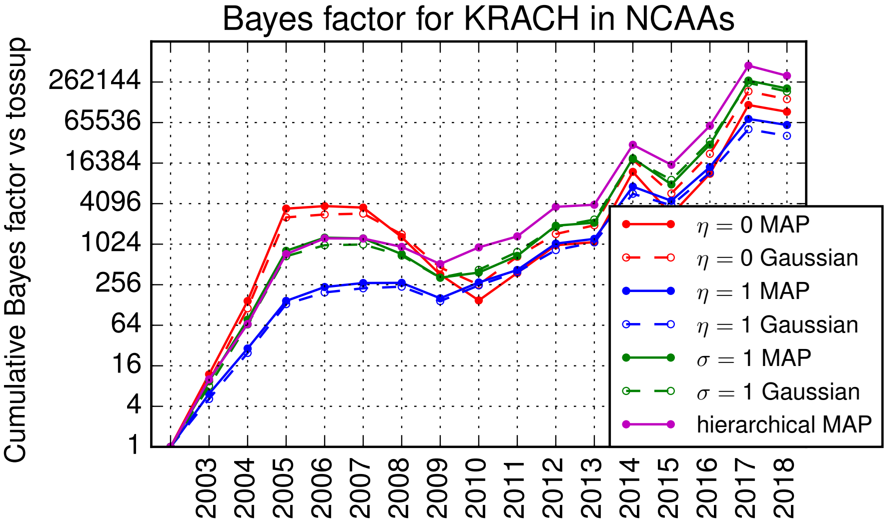
- Gaussian approx w/ $H_{ij} = -n_{ij}\widehat{\theta}_{ij}\widehat{\theta}_{ji} + \delta_{ij}\left(\sum_{k=1}^{t} n_{ik}\widehat{\theta}_{ik}\widehat{\theta}_{ki} + \frac{1}{\widehat{\sigma}^2}\right)$

and $\quad H_{i\,\ln\sigma} = -\frac{\widehat{\lambda}_i}{\widehat{\sigma}^2} \quad$ and $\quad H_{\ln\sigma\,\ln\sigma} = 2\frac{\sum_{i=1}^{t}\widehat{\lambda}_i{}^2}{\widehat{\sigma}^2} + \frac{1}{\varepsilon^2}$

MAP Estimate of Hyperparameter



Cumulative Bayes Factor, 2003-2018



Takeaways

- Bayes factors: evaluate probabalistic predictions after the fact
- Bradley-Terry is definitely better than guessing!
- Hard to distinguish details w/15×16 results
- Hierarchical modelling is harder than it looks
  (but maybe worth it)

References

- Bradley and Terry 1952 *Biometrika* **39**, 324
- Butler 1993 Ken's Ratings for American College Hockey
  http://lists.maine.edu/cgi/wa?A2=Hockey-L;68c89935.9307
- Butler and Whelan 2000 arXiv:math.ST/0412232
- Ford 1957 *American Mathematical Monthly* **64**, 28
- Haldane 1932 *Math. Proc. Cambridge Phil. Soc.* **28**, 1469
- Leonard 1977 *Biometrics* **33**, 121
- Phelan and Whelan 2018 arXiv:1712.05879
- Whelan 2018 arXiv:1712.05311
- Zermelo 1929 *Mathematische Zeitschrift* **29**, 436

EXTRA SLIDES

Finiteness of Maximum Likelihood Ratios

Solutions to $v_i = \sum_{j=1}^{t} w_{ij} = \sum_{j=1}^{t} n_{ij} \frac{\widehat{\pi}_i}{\widehat{\pi}_i + \widehat{\pi}_j} = \sum_{j=1}^{t} n_{ij} \widehat{\theta}_{ij}$

- Obv, if k undefeated ($v_k = \sum_{j=1}^{t} n_{ik}$), $\widehat{\pi}_k \to \infty$ & $\forall j : \widehat{\theta}_{kj} = 1$
- More generally (Albert & Anderson 1984, Santner & Duffy 1986), if you can make a "chain of wins" from i to j, write $i \trianglerighteq j$
  - If $i \trianglerighteq j$ but $j \ntrianglerighteq i$, then $\widehat{\pi}_i / \widehat{\pi}_j \to \infty$, and $\widehat{\theta}_{ij} = 1$
  - If $i \trianglerighteq j$ and $j \trianglerighteq i$, then $\widehat{\pi}_i / \widehat{\pi}_j$ finite, and $0 < \widehat{\theta}_{ij} < 1$
  - If $i \ntrianglerighteq j$ and $j \ntrianglerighteq i$, then $\widehat{\pi}_i / \widehat{\pi}_j$ & $\widehat{\theta}_{ij}$ undetermined
- Butler & Whelan 2000: teams split into "groups" (equiv classes) within which ML ratios are finite;
  ML ratios between groups are 0, ∞ or undefined
  Can summarize in Directed Acyclic Graph

Problems with Maximum Likelihood Estimates

- Infinite or undetermined MLE ratios problematic
- Problem goes away given enough data,
  but compromises use of Bradley-Terry to rank teams
  after a short season (e.g., College Football)
- Counterintuitive:
  - beating an "infinitely worse" team does nothing to MLE
  - impossible to be better than an undefeated team
- Can resort to ad hoc regularization e.g., "fictitious games"
  to force ratios to be finite (Butler, unpublished)
- Motivates a Bayesian approach with prior information
  (at least "nobody's perfect")

Parametrization for Bradley-Terry

π_i	$\lambda_i = \ln \pi_i$	$\zeta_i = \frac{\pi_i}{1+\pi_i}$
$\theta_{ij} = \frac{\pi_i}{\pi_i+\pi_j}$	$\gamma_{ij} = \ln \frac{\theta_{ij}}{1-\theta_{ij}} = \lambda_i - \lambda_j$	$\lambda_i = \ln \frac{\zeta_i}{1-\zeta_i}$

- $p(D|\boldsymbol{\lambda}) = p(D|\boldsymbol{\pi})$ with $\pi_i = e^{\lambda_i}$ but $f(\boldsymbol{\lambda}|X) = e^{\sum_{i=1}^{t} \lambda_i} f(\boldsymbol{\pi}|X)$
- Work with $\boldsymbol{\lambda}$ because $\sum_{i=1}^{t} \frac{\partial \theta_{jk}}{\partial \lambda_i} = 0$, i.e., probabilities $\{\theta_{ij}\}$ depend on combinations "orthogonal" to $\sum_{i=1}^{t} \lambda_i$
- Note if prior $f(\boldsymbol{\lambda}|I)$ is uniform, posterior $f(\boldsymbol{\lambda}|D, I)$ is maximized by maximum likelihood solution $\widehat{\boldsymbol{\lambda}}$